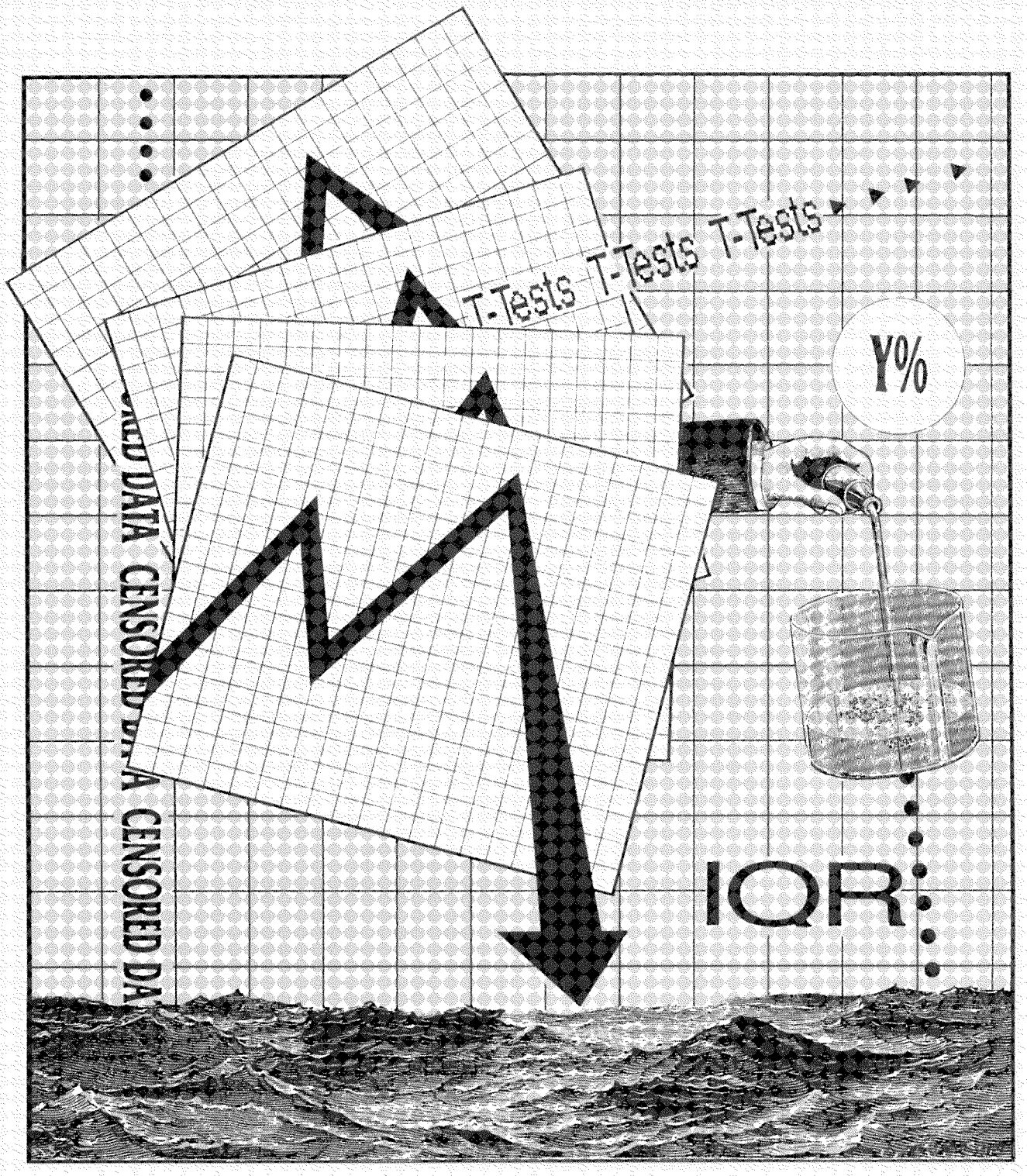
ES&T
FEATURES

Less than obvious

Statistical treatment of data below the detection limit



Dennis R. Helsel
U.S. Geological Survey
Reston, VA 22092

As researchers increasingly investigate trace substances in the world's soil, air, and water, they frequently find concentrations that are lower than limits deemed reliable enough to report as numerical values. These so-called "less-than" values—values stated only as "<*rl*," where *rl* is the "reporting limit" or "limit of quantitation" *(1)* or "determination limit" *(2)*—present a serious interpretation problem for data analysts. For example, compliance with wastewater discharge regulations usually is judged by comparing the mean of concentrations observed over some time interval with a legal standard. Yet mean values from samples cannot be computed when less-thans are present.

Studies of groundwater quality at waste-disposal sites commonly involve comparisons of two groups of data (up-gradient versus down-gradient wells). Usually, *t* tests (the most common test for determining whether two means differ) are employed for this purpose, but the *t* test requires estimates of means and standard deviations that are impossible to obtain unless numerical values are fabricated to replace any less-thans present in the data. The results of such tests can vary greatly depending on the values fabricated. Therefore, estimates of summary statistics (such as mean, standard deviation, median, and interquartile range) that best represent the entire distribution of data, below and above the reporting limit, are necessary to analyze environmental conditions accurately. Also needed are hypothesis test procedures that provide valid conclusions as to whether differences exist among one or more groups of data. These needs must be met using the only information available to the data analyst: concentrations measured above one or more reporting limits, and the observed frequency of data below those limits.

This paper discusses the most appropriate statistical procedures for handling data that have been reported as less-thans. It does not consider the alternative of reporting numerical values for all data, including those below reporting limits *(3–6)*.

Estimating summary statistics

Methods for estimating summary statistics of data that include less-thans (statisticians call these "censored data") can be divided into three classes: simple substitution, distributional, and robust methods. Recent papers have documented the relative performance of these methods *(7–11)*. The first three papers compare the abilities of several estimation methods in detail over thousands of simulated data sets *(7–9)*. They are applied to numerous water-quality data sets, including those that are not similar to the assumed distributions required by the distributional methods *(10)*. A single case study is reported *(11)*. Only one report deals with censoring at multiple reporting limits *(9)*. Large differences in these methods' abilities to estimate summary statistics have been found.

Which summary statistics are appropriate? Environmental quality data usually are positively skewed, and sometimes very highly skewed *(7, 12–14)*. This is especially true for data close to zero that include censored values, because the lower bound of zero ensures a positive skew. In a typical pattern, most data have low values, but a few high "outliers" occur. In such cases, the mean and standard deviation are affected strongly by those few observations that show the highest values. The mean and standard deviation may be quite sensitive to the deletion or addition of even one observation, and therefore are poor measures of central value and variability. For positively skewed data, the mean may be exceeded by less than half of the observations, sometimes even by 25% or less. The mean, therefore, is not a good estimate of the central value of those data. Similarly, the standard deviation will be inflated by outliers, implying a variability larger than that shown by the majority of the data set. The mean and standard deviation are useful for mass loadings of a constituent, such as computations of the average sediment concentration at a river cross section. Large concentrations at one point in the cross section *should* increase the overall mean value. However, when the strong influence of one large value distorts summaries of data characteristics, such as the "typical" sediment characteristics found over many streams, the mean and standard deviation usually are not appropriate measures.

Alternative measures of central value and variability for skewed data are percentile parameters such as the median and interquartile range (IQR). By definition, the median has 50% of the values of the data above it and 50% below. Unlike the mean, the median is not strongly affected by a few low or high "outlier observations." It is a more stable (or "resistant") estimator of typical value for skewed data and is similar to the mean for symmetric (nonskewed) data. Often, the "geometric mean," the mean of logarithms of the data, is computed for the same purpose. The geometric mean is an estimate of the median (in original units) when the logarithms are symmetric.

Like the median, the IQR is largely unaffected by the lowest or highest data values. It is the 75th percentile minus the 25th percentile, and thus is the range of the central 50% of the data. The IQR equals 1.35 times the standard deviation for a normal distribution. However, for the skewed distributions common to environmental monitoring data, the IQR often will be much smaller than the standard deviation, and a better estimate of variability of the bulk of the data.

The median and the IQR have another advantage when applied to censored data: When the values of less than 50% of the data are below the reporting limit, the sample median is known. Similarly, when less than 25% of the data are censored, the sample IQR is known. No "fix-ups" are necessary to obtain sample estimates.

Comparing estimation methods. Estimation methods may be compared on the basis of their ability to replicate true population statistics. Departures from true values are measured by root mean squared error (RMSE), which combines bias and lack of precision. Methods with lower RMSE are considered better.

Class 1: Simple substitution methods. These methods substitute a single value such as one-half the reporting limit for each less-than value. Summary statistics are calculated using these fabricated numbers together with the values above the reporting limit. These methods are widely used, but have no theoretical basis. As Figure 1 shows, the distributions resulting from simple substitution methods have large gaps and do not appear realistic.

All of the studies cited above determined that simple substitution methods perform poorly in comparison with other procedures *(7–11)*. The substitution of zero produces estimates of mean and median that are biased low, whereas substituting the reporting limit results in estimates above the true value. Results for the standard deviation and IQR, and for substituting one-half the reporting limit, also are far less desirable than those for alternative methods. With the advent of convenient software *(11)* for other procedures, there appears to be no reason to use simple substitutions for such computations. Because large differences may occur in the resulting estimates, and as the choice of value for substitution essentially is arbitrary without some knowledge of instrument readings below the reporting limit, estimates resulting from simple substitution are not defensible.

Class 2: Distributional methods. Distributional methods (Figure 2) use the characteristics of an assumed distribution to estimate summary statistics. Values of data below and above the report-

FIGURE 1

Histograms for simple substitution methods for handling less-than values

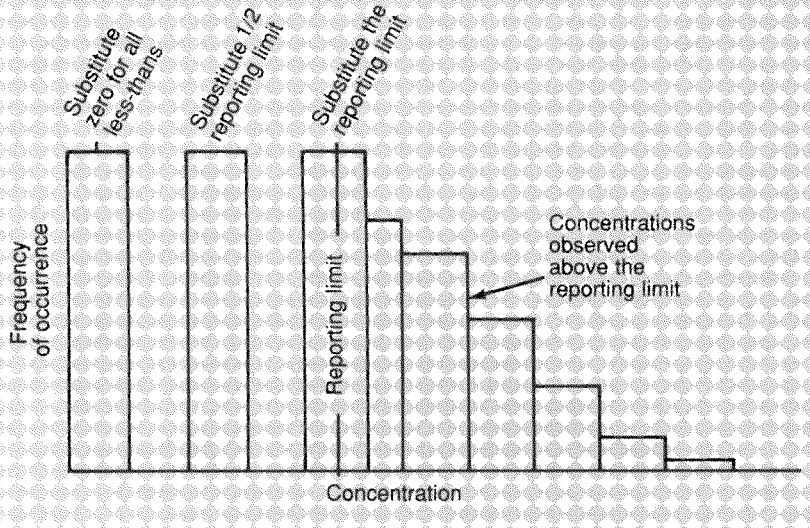
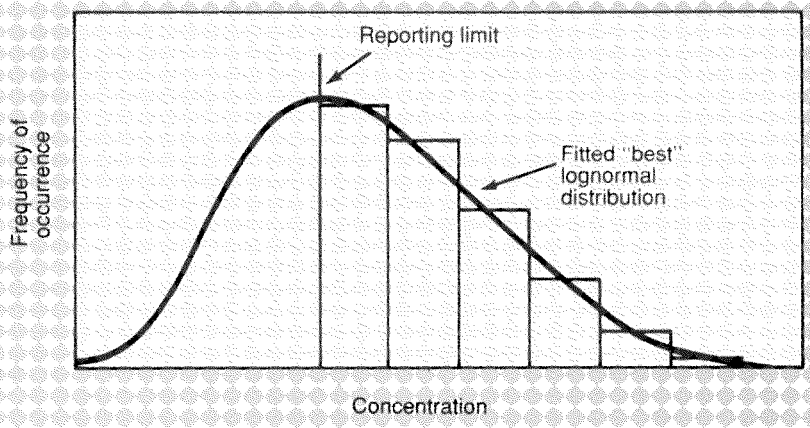


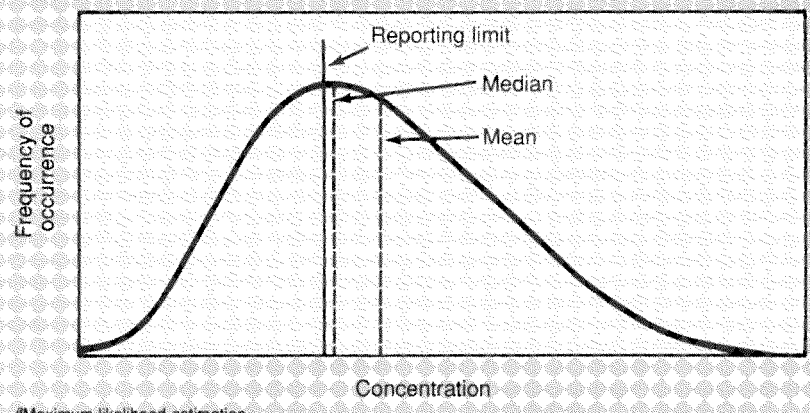
FIGURE 2

Distributional (MLE[a]) method for computing summary statistics

MLE fits "best" lognormal distribution to the data . . .



. . . and then determines summary statistics of the fitted distribution to represent the data



[a]Maximum likelihood estimation.

ing limit are assumed to follow a distribution such as the lognormal. Given a distribution, estimates of summary statistics are computed that best match the observed concentrations above the reporting limit and the percentage of data below the limit. Estimation methods include maximum likelihood estimation (MLE) *(15)* and probability plotting procedures *(16)*. Although MLE estimates are more precise than probability plotting, both methods are unbiased when observations fit the assumed distribution exactly and the sample size is large. This is rarely the case, however. When data do not match the observed distribution, both methods may produce biased and imprecise estimates *(7, 9)*. The most crucial consideration when using distributional methods, then, is how well the data can be expected to fit the assumed distribution. Even when distributional assumptions are correct, MLEs have been shown to produce estimates with large bias and poor precision for the small sample sizes (n = 5, 10, and 15) considered common for environmental data *(8)*. MLE methods commonly are used in environmental disciplines such as air quality studies *(17)* and geochemistry *(12)*.

Assuming a lognormal distribution for concentrations, MLEs for larger data sets (n = 25, 50) have provided excellent estimates of percentiles (median and IQR) for a variety of data distributions realistic for environmental studies, including those that are not lognormal. However, they have not worked as well for estimating the mean and standard deviation *(7, 10)*. There are two reasons this is so.

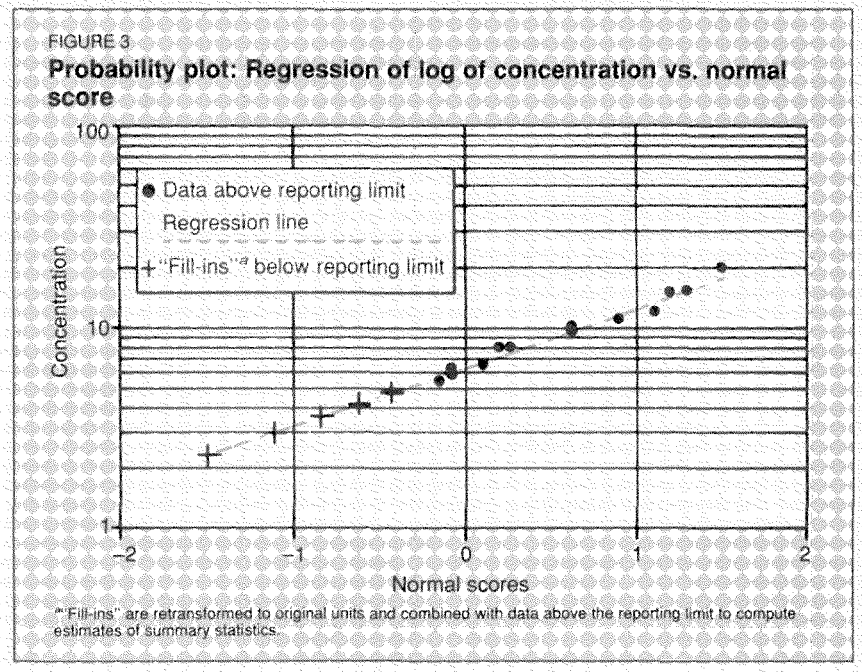
First, the lognormal distribution is flexible in shape and provides reasonable approximations to data which are nearly symmetric, as well as to some positively skewed distributions which are not lognormal. Thus the lognormal can mimic the actual shape of the data over much of the distribution, adequately reproducing percentile statistics even though the data were not truly lognormal in shape. However, the moment statistics (mean and standard deviation) are very sensitive to values of the largest observations. Failure of the assumed distribution to fit these observations will result in poor estimates of moments.

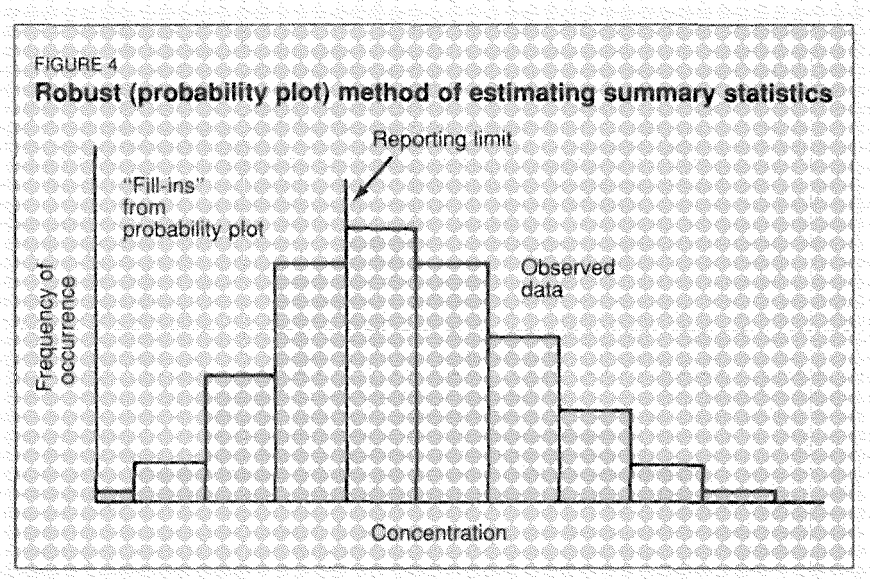
Second, there is a transformation bias (Table 1) inherent in computing estimates of the mean and standard deviation for any transformation—including logarithms—and then transforming back to original units *(18–19)*. Percentiles, however, can be transformed directly between measurement scales without bias. Estimates of mean and standard deviation computed in transformed units by MLEs or other methods are biased

TABLE 1
A simple example of transformation bias

Original data	Base 10 logarithms
1	0
10	1
100	2
1000	3
10,000	4
Mean 2222.2	2^a

[a] Mean of logarithms retransformed = 10^2 = 100 ≠ 2222.2





when they are retransformed. Several studies have included methods that attempt to correct for this bias *(9, 11, 12)*.

Two distributional methods that are used less frequently are a "fill-in with expected values" MLE technique *(8)* and a probability plot method which estimates the mean and standard deviation by the intercept and slope, respectively, of a line fit to data above the reporting limit *(16)*. Probability plot methods are easy to compute with standard statistics software, an advantage for practitioners. Both methods suffer from transformation bias, however, when estimates are computed in one scale and then retransformed back into original units. Therefore, the probability plot has been recommended for estimating the geometric mean *(16)*, but it would not work well for estimating the mean in original units because of transformation bias. Both methods should be slightly less precise than MLEs.

Class 3: Robust methods. These methods (Figure 3) combine observed data above the reporting limit with below-limit values extrapolated, assuming a distributional shape, in order to compute estimates of summary statistics (Figure 4). A distribution is fit to the data above the reporting limit by either MLE or probability plot procedures *(7, 9)*, but the fitted distribution is used only to extrapolate a collection of values below the reporting limit. These extrapolated values are not considered estimates for specific samples, but are used collectively only to estimate summary statistics. The robustness of these methods results primarily from their use of observed data rather than a fitted distribution above the reporting limit. They also avoid transformation bias by performing all computations of summary statistics in original units.

Robust methods have produced consistently small errors for all four summary statistics in simulation studies *(7, 9)*, as well as when applied to actual data *(10)*. Robust methods have at least two advantages over distributional methods for computation of means and standard deviations. First, they are not as sensitive to the fit of a distribution for the largest observations because actual observed data are used rather than a fitted distribution above the reporting limit. Second, estimates of extrapolated values can be directly retransformed and summary statistics computed in the original units, thereby avoiding transformation bias.

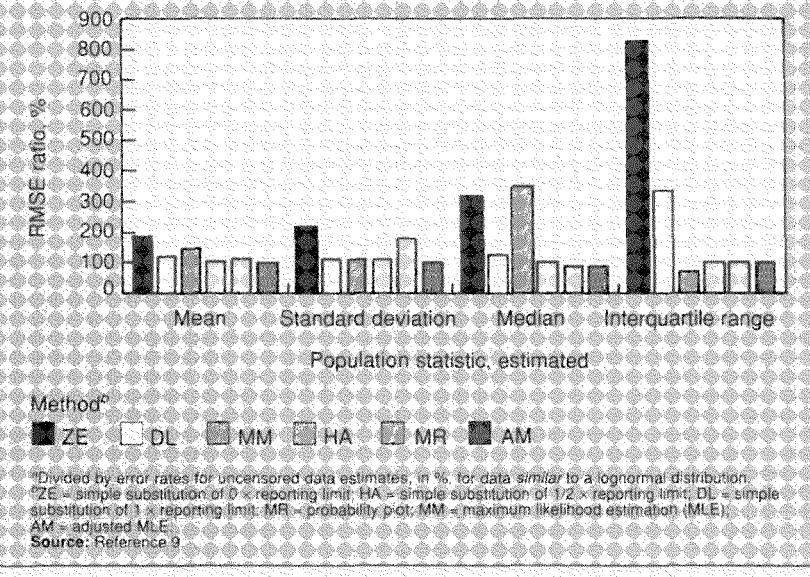
Recommendations. Robust procedures have substantial advantages over distributional methods when concentrations cannot be assumed to follow a defined distribution. In practice, the distribution of environmental data is rarely if ever known, and it may vary between constituents, time periods, and locations. It is not surprising, therefore, that robust methods have been recommended for estimating the mean and standard deviation *(7, 9)*. Either robust probability plot or distributional MLE procedures perform well for estimating the median and IQR *(7–9)*. The use of these methods, rather than simple substitution methods for environmental data, should reduce estimation errors for summary statistics substantially.

Multiple reporting limits. Data sets may contain values censored at more than one reporting limit. This occurs frequently as limits are lowered over time at a single laboratory, or when data having different reporting limits are combined from multiple laboratories. Estimation methods that belong to the three classes described above are available to remedy this situation. A comparison of these methods *(9)* again leads to the conclusion that robust methods provide the best estimates of mean and standard deviation, and MLEs for percentiles. For example, in Figure 5, the error rates for six estimation methods are compared with the error that would occur had all

data been above the reporting limit (shown as the 100% line). Figure 6 shows the same information when the data differ markedly from a lognormal distribution *(9)*. The simple substitution methods ZE, HA, and DL (substitution of zero, one-half, and one times the reporting limit) have more error in most cases than does MR (the robust probability plot method). A lower RMSE that results from the use of substitution methods is an artifact of constant, strongly biased estimates, also not a desirable result. MM (the maximum likelihood procedure) and AM (the MLE adjusted for transformation bias) show themselves to be excellent estimation methods for percentiles, but they suffer from large errors when the mean and standard deviation are estimated. In summary, the use of MLE for estimation of percentiles and of the robust probability plot method for estimating the mean and standard deviation should decrease errors far more efficiently than would simple substitution methods for data with multiple reporting limits.

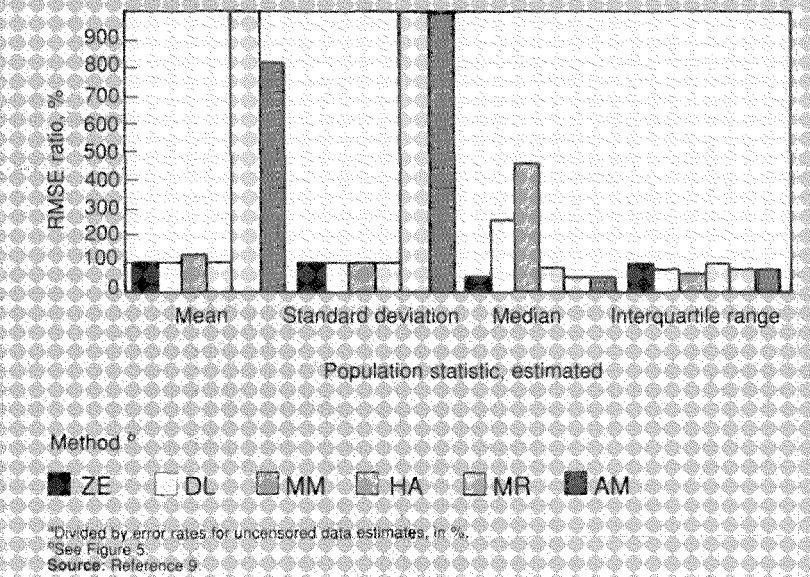
Software for computations. MLE methods require advanced computational software. These and other distributional methods for single reporting limits, including the distributional (slope-intercept) probability plot estimator, recently were made available to the scientific community *(11)*. By contrast, the robust probability plotting method for a single reporting limit can be computed easily by most commercially available statistics software. Normal scores ("NSCORES" of Minitab, or "PROC RANK" within SAS, for example) first are computed with all less-thans set to slightly different values all below the reporting limit. Second, a linear regression equation is developed using only the above-limit observations, where log of concentration is the y variable and normal scores the x variable. Estimates for the below-limit data then are extrapolated using this regression equation from normal scores for the below-limit data. Finally, extrapolated estimates are retransformed into units of concentration, combined with above-limit concentration data, and summary statistics computed. Fortran code for multiple reporting limit techniques may be obtained by sending a self-addressed, stamped envelope and a formatted 3 1/2 inch disk (MS-DOS or Macintosh format) to the author.

FIGURE 5
Root mean square error rates of six multiple-detection methods for data similar to a lognormal distribution[a]



[a]Divided by error rates for uncensored data estimates, in %, for data *similar* to a lognormal distribution. [b]ZE = simple substitution of 0 × reporting limit; HA = simple substitution of 1/2 × reporting limit; DL = simple substitution of 1 × reporting limit; MR = probability plot; MM = maximum likelihood estimation (MLE); AM = adjusted MLE.
Source: Reference 9

FIGURE 6
Root mean square error rates of six multiple-detection methods for data *not* similar to a lognormal distribution[a]



[a]Divided by error rates for uncensored data estimates, in %. [b]See Figure 5.
Source: Reference 9

Methods for hypothesis testing

Methods for hypothesis testing of censored data can be classified into the three types: simple substitution (class 1), distributional or parametric (class 2), and robust or nonparametric (class 3). Parametric statistical tests frequently are used in environmental assessments. They assume that data follow some distributional shape, usually the normal distribution. Parameters (true population statistics), such as the mean and standard deviation, are estimated to perform the test. When censoring is present, values often are fabricated to estimate these parameters (class 1). Problems caused by fabrication are illustrated below. Parametric tests that do not require substitutions for less-thans (class 2) also are available. Where the distributional assumptions are appropriate, these relatively unknown tests are very useful.

Investigators have, on occasion, deleted censored data before hypothesis testing. This approach is the worst procedure because it causes a large and variable bias in the parameter estimates for each group. After deletion, comparisons made are between the upper X % of one group versus the upper Y % of another, where X and Y may be very different. Such tests have little or no meaning.

Alternatively, nonparametric tests can be performed *(20)*. These tests simply rank the data and indicate whether the ordering of the data points shows that

differences occur or that trends exist. No fabrication of data values is required because all censored data are represented by ranks that are tied at values lower than the lowest number above the reporting limit. These tests generally have greater power than parametric tests when the data do not conform to a normal distribution *(20, 21)*.

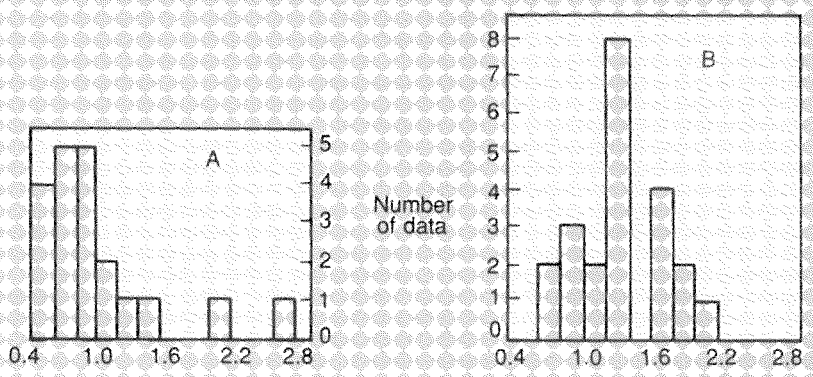
As an example of the differences between hypothesis test methods for censored data, tests were performed that determine whether means or medians significantly differ between two groups. Two data sets were generated from lognormal distributions having the same variance but differing mean values. Sample statistics for the two data sets before and after censoring are given in the box.

Before any censoring, group means are shown to be significantly different by a *t* test ($p = 0.04$, Table 2) and by a *t* test for regression slope equal to zero. The latter is performed by designating the data set each observation belongs to as either a zero or one. This binary variable then is used as the explanatory (independent) variable in a linear regression. Though identical to the *t* test before censoring, a variation of the regression approach will become the distributional (class 2) method for censored data used later. The equivalent nonparametric test, the rank-sum test, produces a much lower *p*-value ($p = 0.003$). This lower *p*-value is consistent with the proven greater power of the nonparametric test to detect differences between groups of skewed data *(21, 22)*, compared with the *t* test.

Suppose that these data represent dissolved arsenic concentrations. A typical reporting limit for dissolved arsenic is 1 μg/L; therefore all data below 1.0 would be recorded as <1. Censoring these data sets at 1 produces 14 less-than values (70%) in group A and five less-than values (23%) in group B (box).

The class 1 method for comparing two groups of censored data is to fabricate data for all less-than values, and include these "data" with detected observations when performing a *t* test. No a priori arguments for fabrication of any particular value between zero and the reporting limit can be made. When zero is substituted for all less-than values, the means are declared significantly different ($p = 0.01$). Yet when the reporting limit of 1.0 is substituted, the means are not found to be different ($p = 0.19$). The conclusion therefore is strongly dependent on the value substituted! This example shows that the fabrication of data followed by a *t* test must be considered too arbitrary for use, especially for legal or management decision purposes, and should be avoided.

Characteristics of two lognormal data groups



	A	B
Number of observations	20	22
Mean	1.00	1.32
Standard deviation	0.57	0.39
25th percentile	0.65	1.07
Median	0.85	1.25
75th percentile	1.07	1.66
Number < reporting limit	14	5

TABLE 2
Example of significance tests between lognormal data groups A and B

Hypothesis test used	Test statistic	*p*
Uncensored data		
t test (Satterthwaite approx.)	−2.13	0.04
Regression with binary variable	−2.17	0.04
Rank-sum test	−2.92	0.003
After imposing artificial reporting limit		
t test		
Less-thans = 0.0	−2.68	0.01
Less-thans = 0.5	−2.28	0.03
Less-thans = 1.0	−1.34	0.19
Tobit regression with binary variable	−2.28	0.03
Rank-sum test	−3.07	0.002

The distributional (class 2) method for hypothesis testing also requires an assumption of normality, but does not involve the substitution of values for censored data. Instead, a *t* test is performed using a regression procedure for censored data known as *tobit regression (23, 24)*. Tobit regression uses the data values above the reporting limit and the proportion of data below the reporting limit to compute a slope coefficient by maximum likelihood. For a two-group test, the explanatory variable in the regression equation is the binary variable of group number, so that data in one group have a value of zero, and in the other group a value of one. The regression slope then equals the difference between the two group means, and the *t* test for whether this slope differs from zero also is a test of whether the group means differ. (Tobit regression is also discussed later in the section on regression.) One advantage of tobit regression for hypothesis testing is that multiple reporting limits may be easily incorporated. Users should be cautioned, however, that proper application requires that the data in both groups be normally distributed around their group mean and that the variance in each group be equal. For large amounts of censoring, these restrictions are difficult to verify.

The nonparametric (class 3) equivalent is the rank-sum test. It considers the 19 less-than values tied at the lowest value, with each assigned a rank of 10 (the mean of ranks 1–19). The resulting *p*-value is 0.002, essentially the same as for the original data, and the two groups are easily declared different. In this example, the nonparametric method makes very efficient use of the information contained in the less-than values, avoids arbitrary assignment of fabricated values, and accurately represents the lack

of knowledge below the reporting limit. Results do not depend on a distributional assumption *(25)*.

When severe censoring (near 50% or more) occurs, all of the above tests have little power to detect differences in central values. The investigator will find it difficult to state conclusions about the relative magnitudes of central values, and other characteristics must be compared. For instance, contingency tables (class 3) can test for a difference in the proportion of data above the reporting limit in each group *(20)*. This test can be used when the data are reported only as detected or not detected. It also may be used when response data can be categorized into three or more groups, such as below detection, detected but below some health standard, and exceeding standards. The test determines whether the proportion of data falling into each response category differs as a function of different explanatory groups, such as different sites or land use categories.

Hypothesis testing with multiple reporting limits. More than one reporting limit often is present in environmental data. When this occurs, hypothesis tests such as comparisons between data groups are greatly complicated. The fabrication of data followed by computation of *t* tests or similar parametric procedures is at least as arbitrary with multiple reporting limits as with one reporting limit, and should be avoided. Also, data below all reporting limits should never be deleted before testing.

Tobit regression (class 2) can be used with multiple reporting limits. Data should have a normal distribution around all group means and equal group variances to use the test. These assumptions are difficult to verify with censored data, especially for small data sets.

One robust method that always can be used is to censor all data at the highest reporting limit, and then perform the appropriate nonparametric test. Thus the data set

<1 <1 <1 5 7 8 <10 <10 <10 12 16 25

would become

<10 <10 <10 <10 <10 <10 <10 <10 <10 12 16 25

and a rank-sum test would be performed to compare this with another data set. Clearly, this causes a loss of information which may be severe enough to obscure actual differences between groups (a loss of power). For some situations, however, this is the best that can be done.

Alternatively, nonparametric score tests common in the medical "survival analysis" literature sometimes can be applied to the case of multiple reporting limits *(26)*. These tests modify uncensored rank test statistics to compare groups of data. The modifications allow for the presence of multiple reporting limits. In the most comprehensive review of these score tests *(27)*, most of them were found inappropriate for the case of unequal sample sizes. Another crucial assumption of score tests is that the censoring mechanism must be independent of the effect under investigation (see box). Unfortunately, this often is not the case with environmental data. The Peto–Prentice test with an asymptotic variance estimate was found to be the least sensitive to unequal sample sizes and to differing censoring mechanisms *(27)*.

In summary, robust hypothesis tests have several advantages over their distributional counterparts when they are applied to censored data. These advantages include freedom from adherence to a normal distribution; greater power for the skewed distributions common to environmental data; comparisons between central values such as the median, rather than the mean; and the incorporation of data below the reporting limit without fabrication of values or bias. Information contained in less-than values is used accurately and does not misrepresent the state of that information.

When adherence to a normal distribution can be documented, tobit regression (class 2) offers the ability to incorporate multiple reporting limits regardless of a change in censoring mechanism. Score tests (class 3) require consistency in the censoring mechanism with respect to the effect being tested.

Methods for regression

With censored data, the use of ordinary least squares (OLS) for regression is prohibited. Coefficients for slopes and intercept cannot be computed without values for the censored observations, and substituting fabricated values may produce coefficients strongly dependent on the values substituted. Four alternative methods capable of incorporating censored observations are described below. The first and last approaches, Kendall's robust fit *(28)* and contingency tables *(20)*, are nonparametric (class 3) methods requiring no distributional assumptions. Robust correlation coefficients also are mentioned *(20)*. Tobit and logistic regression *(24, 29)*, the second and third methods, fit lines to data using maximum likelihood (class 2). Both methods assume normality of the residuals, though with logistic regression, the assumption is after a logit

The appropriateness of score tests

When a score test is not appropriate

Score tests are inappropriate when the censoring mechanism differs for the two groups. That is, the probability of obtaining a value below a given reporting limit differs for the two groups when the null hypothesis that the groups are identical is true.

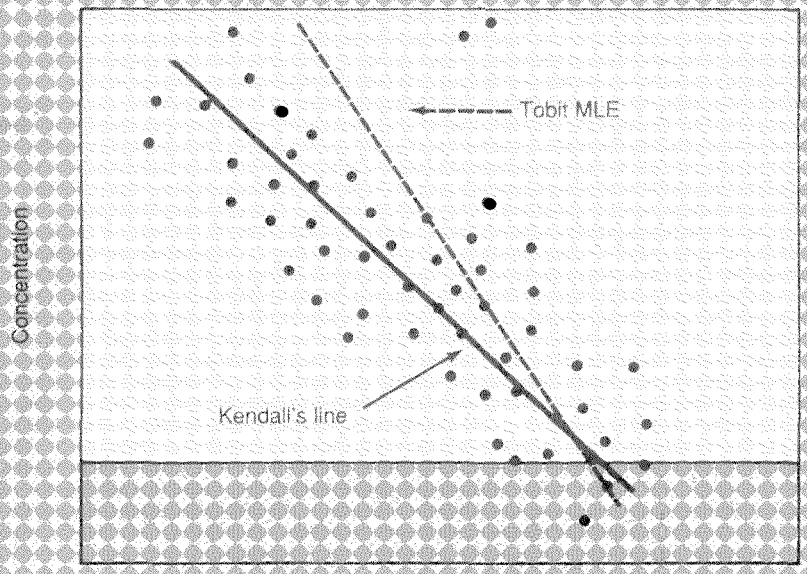
1. Suppose a trend over time is being investigated. The first five years of data are produced by a method that has a reporting limit of 10 µg/L; the second five years of data are compiled by an improved method with 1 µg/L as its reporting limit. A score test of the first half of the data versus the second would not be valid because the censoring mechanism itself varies as a direct function of time.
2. Two groups of data are compared as in a rank-sum test, but most of the data from group A were measured with a chemical method having 1 as its reporting limit, and most of group B were measured with a method having 10 as its reporting limit. A score test would not yield valid results because the censoring mechanism varies as a function of what is being investigated (the two groups).

When a score test is appropriate

A score test yields valid results when the change in censoring mechanism is not related to the effect being measured. Stated another way, the probability of obtaining data below each reporting limit is the same for all groups, assuming that the null hypothesis of no trend or no difference is true. Here a score test provides much greater power than does artificially censoring all data below the highest reporting limit before using the rank-sum test.

1. Comparisons have been made between two groups of data collected at roughly the same time and analyzed by the same methods, even though those methods and reporting limits have changed over time. Score tests are valid in this case.
2. Differing reporting limits result from analyses performed at different laboratories, but each sample had been assigned at random to the different laboratories. Censoring thus is not a function of what is being tested, but is a random effect, and score tests would be valid.

FIGURE 7
Kendall's and tobit MLE[a] lines for censored data with outliers[b]



[a]Maximum likelihood estimation. [b]Note that the tobit line is influenced strongly by outliers.

transformation *(24)*. As before, assumptions are sometimes hard to check with censored data.

The choice of method depends on the amount of censoring present as well as on the purpose of the analysis. For small amounts of censoring (below 20%), either Kendall's line or the tobit line may be used. Kendall's line would be preferred if the residuals are not normally distributed, or when outliers are present. For moderate censoring (20–50%), tobit or logistic regression must be used. With large amounts of censoring, inferences about concentrations themselves must be abandoned, and logistic regression must be employed. When the explanatory and response variables are censored, tobit regression is applicable for small amounts of censoring. For larger amounts of censoring, contingency tables or rank correlation coefficients are the only option.

Kendall's robust line fit. When one censoring level is present, Kendall's rank-based procedure for fitting a straight line to data can test the significance of the relationship between a response and an explanatory variable *(28)*. Also of interest is an equation for the line, including an estimate of the slope. This can be computed when the amount of censoring is small.

Kendall's estimate of slope is the median of all possible pairwise slopes of the data. To compute the slope with censoring, compute the median of all possible slopes twice, once with zero substituted for all less-thans and once with the reporting limit substituted. For small amounts of censoring, the resulting slope will change very little or not at all, and can be reported as a range if necessary. If the slope value change is of an unacceptable magnitude, tobit or logistic regression must be performed. Research currently is underway on methods based on scores that may allow robust regression fits to data with multiple reporting limits *(30)*.

Tobit regression. Censored response data can be incorporated together with uncensored observations into a procedure called tobit regression *(23, 24)*. It is similar to OLS except that the coefficients are fit by maximum likelihood estimation. MLE estimates of slope and intercept are based on the assumption that the residuals are normally distributed around the tobit line, with constant variance across the range of predicted values. Again, it is difficult to check these assumptions with censored data. Outliers can have a strong influence on the location of the line and on significance tests (Figure 7), as is true with uncensored OLS. Residuals for uncensored data should be plotted versus predicted values, so that linearity and constant variance assumptions can be verified for at least small amounts of censoring. For larger percentages of less-thans, decisions whether to transform the response variable often must be made on the basis of previous knowledge (e.g., "metals always need to be log-transformed").

Tobit regression also is applicable when the response and explanatory variables are censored; for instance, in a regression relationship between two chemical constituents. The amount of censoring, however, must be sufficiently small that the linearity, constant variance, and normality assumptions of the procedure can be checked. Cohn *(18)* and others have proven that the tobit estimates are slightly biased and have derived bias corrections for the method.

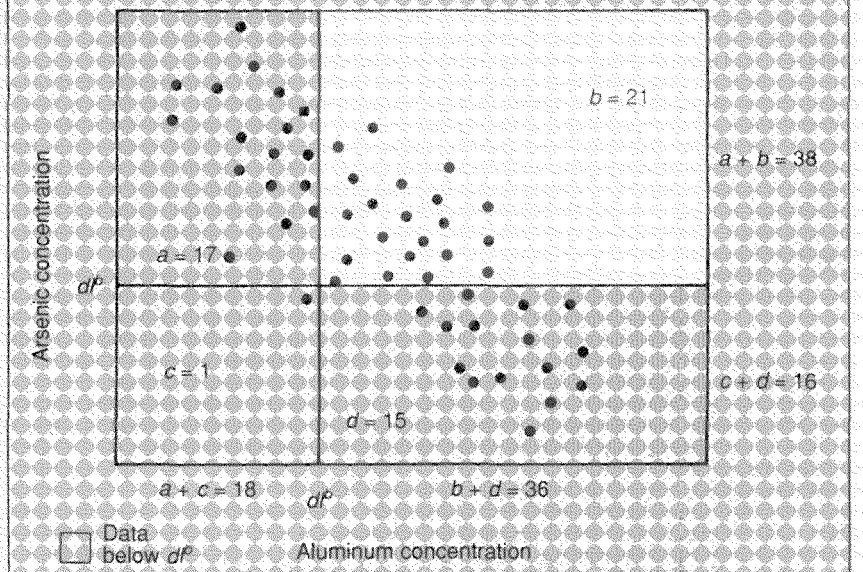
Logistic regression. Here, the response variable is categorical *(29)*. This method does not predict concentration, but rather a probability of being in discrete binary categories such as above or below the reporting limit. A response above the limit usually is assigned a value of 1, and below the limit a 0. The probability of being in one category versus the other is tested to see if it differs as a function of continuous explanatory variable(s). Examples include predicting the probability of detecting concentrations of some organic contaminant from continuous variables such as nitrate concentrations, population density, percent of some appropriate land use variable, or irrigation intensity. Predictions from this regression-type relationship will fall between 0 and 1, and are interpreted as the probability (p) of observing a response of 1. Therefore $[1 - p]$ is the probability of a zero response.

Logistic regression may be used to predict the probabilities of more than two response categories. When there are $m > 2$ ordinal (i.e., may be placed in an order) responses possible, $(m - 1)$ equations must be derived from the data. For example, if three responses are possible (concentrations below rl = 0, above rl but below health standards = 1, and above health standards = 2), two logistic regressions must be computed. First, an equation must be written for the probability of being nonzero (the probability of being above the rl). Next, the probability of a 2 (probability of being above the health standard) also is modeled. Together, these two equations completely define the three probabilities $p\,(y = 0)$, $p\,(y = 1)$, and $p\,(y = 2)$ as a function of the explanatory variables.

Contingency tables. Contingency tables are useful in the regression context if both explanatory and response variables contain censoring *(20)*. For example, suppose the relationship between two trace metals in soils (such as arsenic and aluminum) is to be described. The worst procedure again would be to delete the data below the reporting limits and perform a regression. Figure 8 shows that a true linear relationship with negative slope could be completely obscured if censored data are ignored and only data in the upper right quadrant in-

FIGURE 8

Contingency table relationship between two censored variables[a]



[a]Ignoring censored data would produce the misleading conclusion that no relationship exists between the two variables. [b]Detection limit.

vestigated. Contingency tables provide a measure of the strength of the relationship between censored variables—the phi statistic ϕ *(20)*, a type of correlation coefficient. An equation that describes this relationship, in the context of regression, is not available. Instead, the probability of y being in one category can be stated as a function of the category of x. For the data in Figure 8, the probability of arsenic being above the reporting limit is 21/36 = 0.58 when aluminum is above reporting limit, and 17/18 = 0.94 when aluminum is below the reporting limit.

Rank correlation coefficients. The robust correlation coefficients Kendall's τ or Spearman's ρ *(20)* also could be computed when both variables are censored. All values below the reporting limit for a single variable are assigned tied ranks. Rank correlations do not provide estimates of the probability of exceeding the reporting limit as does a contingency table. Therefore, they are not applicable in a regression context, but would be more applicable than contingency tables in a correlation context. One such context would be in "chemometrics" *(31)*, the computation of correlation coefficients for censored data as inputs to a principal components or factor analysis.

In summary, relationships between variables with data below reporting limits can be investigated in a manner similar to regression. Values should not be fabricated for less-thans before regression. Instead, for small amounts of censoring and one reporting limit, Kendall's robust line can be fit to the data. For moderate censoring or multiple reporting limits, tobit regression can be performed. For more severe censoring of the dependent variable, logistic regression is appropriate. When response and explanatory variables contain severe censoring, contingency tables can be performed.

Less-thans are valuable data

Methods are available that appropriately incorporate data below the reporting limit for purposes of estimation, hypothesis testing, and regression. The deletion of censored data or fabrication of values for less-thans leads to undesirable and unnecessary errors.

Acknowledgments

The author wishes to thank the reviewers for comments which greatly improved the manuscript. Use of trade names does not imply endorsement by the U.S. Geological Survey.

This article was reviewed for suitability as an *ES&T* feature by Richard Gilbert, Battelle Northwest Laboratories, Richland, WA 99352; and P. Steven Porter, Everglades Research and Education Center, Belle Glade, FL 33430.

References

(1) Keith, L. H. et al. *Anal. Chem.* **1983,** *55,* 2210–18.
(2) Currie, L. A. *Anal. Chem.* **1968,** *40,* 586–93.
(3) ASTM Subcommittee D19.02. *Annual Book ASTM Standards 1983, 11.01,* Chapter D; American Society for Testing and Materials: Philadelphia, 1983; pp. 4210–83.
(4) Porter, P. S.; Ward, R. C.; Bell, H. F. *Environ. Sci. Technol.* **1988,** *22,* 856–61.
(5) Gilliom, R.J.; Hirsch, R. M.; Gilroy, E. J. *Environ. Sci. Technol.* **1984,** *18,* 530–35.
(6) Porter, P. S. In *Monitoring to Detect Changes in Water Quality Series;* IAHS Publication no. 157; Lerner, D., Ed.; International Association of Hydrological Sciences: Wallingford, England, 1986; pp. 305–15.
(7) Gilliom, R.; Helsel, D. *Water Resour. Res.* **1986,** *22,* 135–46.
(8) Gleit, A. *Environ. Sci. Technol.* **1985,** *19,* 1201–06.
(9) Helsel, D.; Cohn, T. *Water Resour. Res.* **1988,** *24,* 1997–2004.
(10) Helsel, D.; Gilliom, R. *Water Resour. Res.* **1986,** *22,* 147–55.
(11) Newman, M. C.; Dixon, P. M. *American Environmental Laboratory* **1990,** 26–30.
(12) Miesch, A. U.S. Geological Survey Professional Paper 574-B; U.S. Geological Survey: Reston, VA, 1967.
(13) Davis, G. D. *Ground Water* **1966,** *4(4),* 5–12.
(14) Luna, R. E.; Church, H. W. *J. Appl. Meteorol.* **1974,** *13,* 910–16.
(15) Cohen, A. C. *Technometrics* **1959,** *1,* 217–23.
(16) Travis, C. C.; Land, M. L. *Environ. Sci. Technol.* **1990,** *24,* 961–62.
(17) Owen, W.; DeRouen, T. *Biometrics* **1980,** *36,* 707–19.
(18) Cohn, T. U.S. Geological Survey Open-File Report 88-350; U.S. Geological Survey: Reston, VA, 1988.
(19) Miller, D. M. *American Statistician* **1984,** *38,* 124–26.
(20) Conover, W. *Practical Nonparametric Statistics,* 2nd ed.; Wiley: New York, 1980.
(21) Blair, R. C.; Higgins, J. J. *Journal of Educational Statistics* **1980,** *5,* 309–35.
(22) Hodges, J. L.; Lehmann, E. L. *Annals of Mathematical Statistics* **1956,** *27,* 324–35.
(23) Powell, J. L. *Econometrica* **1986,** *54,* 1435–60.
(24) Judge, G. G. et al. *The Theory and Practice of Econometrics;* Wiley: New York, 1980; chap. 14.
(25) Helsel, D.; Hirsch, R. *Water Resour. Bull.* **1987,** *24,* 201–04.
(26) Millard, S.; Deverel, S. *Water Resour. Res.* **1988,** *24,* 2087–98.
(27) Latta, R. *J. Am. Stat. Assoc.* **1981,** *76,* 713–19.
(28) Lehmann, E. *Nonparametrics: Statistical Methods Based on Ranks;* Holden-Day Publishers: Oakland, CA, 1975.
(29) Amemiya, T. *Journal of Economic Literature* **1981,** *19,* 1483–1536.
(30) McKean, J.; Sievers, G. *Technometrics* **1989,** *31,* 207–18.
(31) *Environmental Applications of Chemometrics;* Breen, J. J.; Robinson, P. E., Eds.; ACS Symposium Series 292; American Chemical Society: Washington, DC, 1985.

Dennis R. Helsel *is chief of the Branch of Systems Analysis, Water Resources Division, U.S. Geological Survey (Reston, VA). He received his Ph.D. in environmental sciences and engineering from Virginia Polytechnic Institute and State University in 1978. Since that time, his research has emphasized the statistical and graphical analysis of surface water and groundwater quality data.*